Published online 23 October 2012 *Nucleic Acids Research, 2013, Vol. 41, Database issue* ***D571–D578***
doi:10.1093/nar/gks984

The Papillomavirus Episteme: a central resource for papillomavirus sequence data and analysis

Koenraad Van Doorslaer[1], Qina Tan[2], Sandhya Xirasagar[2], Sandya Bandaru[2], Vivek Gopalan[2], Yasmin Mohamoud[2], Yentram Huyen[2] and Alison A. McBride[1,*]

[1]DNA Tumor Virus Section, Laboratory of Viral Diseases and [2]Bioinformatics and Computational Biosciences Branch, Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD 209892, USA

Received August 15, 2012; Revised September 26, 2012; Accepted September 27, 2012

ABSTRACT

The goal of the Papillomavirus Episteme (PaVE) is to provide an integrated resource for the analysis of papillomavirus (PV) genome sequences and related information. The PaVE is a freely accessible, web-based tool (http://pave.niaid.nih.gov) created around a relational database, which enables storage, analysis and exchange of sequence information. From a design perspective, the PaVE adopts an Open Source software approach and stresses the integration and reuse of existing tools. Reference PV genome sequences have been extracted from publicly available databases and reannotated using a custom-created tool. To date, the PaVE contains 241 annotated PV genomes, 2245 genes and regions, 2004 protein sequences and 47 protein structures, which users can explore, analyze or download. The PaVE provides scientists with the data and tools needed to accelerate scientific progress for the study and treatment of diseases caused by PVs.

INTRODUCTION

The Papillomaviridae (papillomaviruses; PVs) are a diverse family of small, double-stranded, circular viruses that infect many (if not all) amniotes. Despite a relatively small genomic size (~7000–8000 bp) and very similar genomic organization, PVs exhibit a highly complex and diverse biology. Individual PVs have traditionally been designated 'viral types'. The full-length genome of these PV types has to be cloned, and its L1 nucleotide sequence cannot share >10% sequence similarity to any other (known) PV type (1,2). Each virus type is species-specific and trophic for a specific anatomical region of the cutaneous or mucosal epithelium. Infections caused by many PVs are sub-clinical (3) while other PV infections manifest as benign tumors known as warts or papillomas. This results in a spectrum of disease ranging from a variety of prolific benign lesions such as common warts, anogenital warts and laryngeal papillomatosis to malignant carcinomas. Importantly, in humans, persistent infection with specific subsets of high-oncogenic risk PVs is responsible for virtually all cervical cancer and for a subset of head and neck cancers (4).

Most PVs encode about eight proteins. Only four proteins are highly conserved and are encoded by all PVs; the E1 and E2 regulatory proteins modulate viral transcription and replication and the L1 and L2 structural proteins compose the viral capsid (for a review, see (5)). In addition, three small and less well-conserved proteins, E5, E6 and E7, play important roles in cell growth regulation and immune evasion. In the cancer-associated PVs, the E6 and E7 proteins function as oncogenes (6). The gene for the E4 protein overlaps the E2 open reading frame (ORF) and encodes a divergent and very abundant late protein. *Cis*-responsive elements required for transcription and replication are located in a region designated as the upstream regulatory region (URR) or long control region (Figure 1).

The relatively small genomic size of PVs, coupled with recent advances in cloning and sequencing methods, have made it possible to obtain the complete genomic sequences of most identified viral types. Papillomaviridae are genetically stable viruses with a slow evolutionary rate (7) and sequence analysis has illustrated that recombination has had only a minor impact on viral evolutionary history (8). Thus, genetic differences between viral types are most likely the result of genetic drift (9). The simple evolutionary history of PVs, the wealth of sequence information as well as the abundant epidemiological and biochemical data mean that the comparative genomics of PV genomes has the potential to elucidate many aspects of the viral life cycle.

Thousands of PV-related sequences are available from public databases. A quick GenBank search for PV (nucleotide database: 'Papillomavirus'[Organism]; accessed 26 July 2012) returned 16 254 related sequences. In addition

*To whom correspondence should be addressed. Tel: +301 496 1370; Fax: +301 451 5330; Email: amcbride@nih.gov

Published by Oxford University Press 2012.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by-nc/3.0/), which permits non-commercial reuse, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com.

Downloaded from http://nar.oxfordjournals.org/ by guest on December 9, 2015

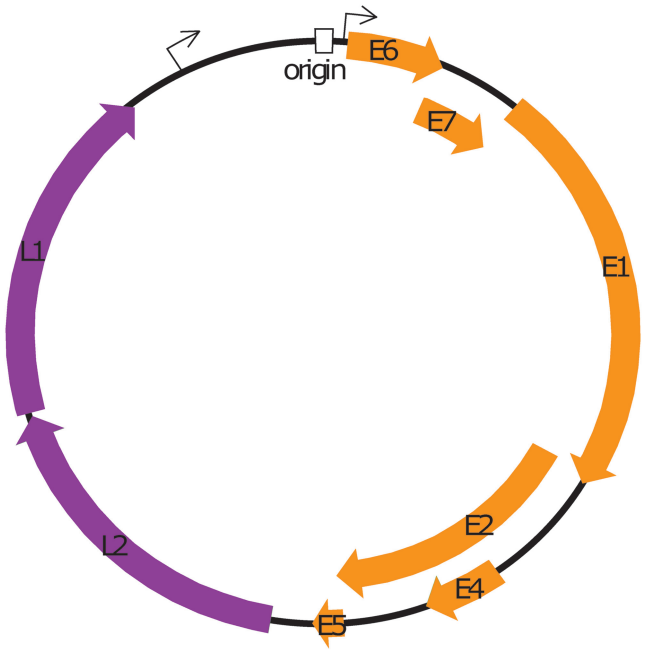

Figure 1. Papillomavirus Genomic Organization. The double-stranded, circular genomes of PVs are ~7–8 kb in length. All PVs encode the E1 and E2 replication proteins and the L1 and L2 capsid proteins. In addition to these core proteins, most viruses encode auxiliary proteins E5, E6 and E7, which manipulate host cell proliferation and cell cycle checkpoints. The URR is located between the L1 and E6 ORFs. This region contains viral promoters, enhancers and the replication origin.

to the overwhelming amount of available sequence data in these databases, genomic annotation is either absent or non-uniform, which further complicates efforts to obtain and compare prototype genomes and gene products. The motto 'Garbage in, garbage out' is a key consideration for successful bioinformatics analyses. However, a central repository of uniformly annotated and community approved prototype genomes has not been available since a database, developed and hosted by the Los Alamos National Laboratory, ceased to be updated in 1997. The main goal of the Papillomavirus Episteme (PaVE) project is to provide researchers with corrected and uniformly annotated Reference genomes. In addition to providing intuitive ways to access and download this sequence information, the PaVE hosts an array of web-based tools to facilitate in the analyses of these sequences.

Episteme is derived from the Greek word for knowledge or science and is defined as 'the body of ideas that determine the knowledge that is intellectually certain at any particular time'. This term embodies our goal to provide highly organized and curated PV sequence information and tools for the research community to accelerate scientific progress and ultimately our understanding, detection, diagnosis and treatment of diseases caused by PVs.

DESCRIPTION

Software backbone

The PaVE utilizes open platform web technologies and is deployed on a server running Apache Tomcat 7 and a MySQL database (v5.1). The underlying code is written in Java 1.6 and uses Hibernate 3.3.1 mapping for BioJava 1.8. The web graphical user interface is developed using the Google Web Toolkit v.2.0.4.

PV sequence-related data

As of July 2012, the PaVE contains integrated information from 241 completely annotated viral genomes. Except for the Los Alamos corrected sequences described below, the original viral genome sequences were acquired from the National Center for Biotechnology Information (NCBI) GenBank. NCBI Reference Sequences (10) were used when available. Experimentally determined structures of PV proteins were obtained from the Protein Data Bank (PDB; www.pdb.org).

Annotation of viral genomes

The process used to annotate viral genomes is outlined in Figure 2. For the initial step, all-possible ORFs were identified on the forward strand of each complete viral genome. To find and annotate each viral gene, the translated ORFs were searched against a custom database consisting of representative viral protein sequences using BLASTp (11). The custom BLAST database contains manually annotated proteins from viruses representing the three main clades in the PV phylogeny (i.e. BPV1, HPV16 and HPV8). This computational step was followed by manual quality control. During this manual step, the annotation was checked to confirm the presence of all expected ORFs. These expectations were based on phylogenetic classification and PV biology. For example, all PVs should encode E1, E2, L1 and L2 ORFs and most PVs also encode an E6 and E7 proteins, but several clades in the phylogenetic tree lack these ORFs (1). Manual annotation was attempted when the automated pipeline missed an expected ORF and, if successful, the newly identified ORF was added to the PaVE database as well as to the custom annotation BLAST database to improve the sensitivity of the pipeline. In some cases (Table 1), viral genome sequences were edited to restore a viral ORF (see below). Upon completion of these steps, a multiple sequence alignment of all homologous viral ORFs was constructed (MAFFT algorithm; (12)). This alignment was used in a final quality control step to verify that the proteins were of the expected length and that no frame shifts or deletions were present within the ORF. Viral ORFs were only added to the PaVE database when they had successfully completed the pipeline and manual quality control. The latest release of the database contains in addition proteins such as E1^E4, which is encoded by a spliced message (Figure 3). Future updates will include additional well-characterized proteins that are the result of splice events.

In addition to viral ORFs, the PaVE contains annotation for *cis*-regulatory elements, such as DNA-binding sites for the viral E2 regulatory protein. Future updates will include uniform annotation for additional elements such as promoters, poly-A addition sites and replication elements. However, where available from the source NCBI files, all of this information is already readily accessible in

Downloaded from http://nar.oxfordjournals.org/ by guest on December 9, 2015

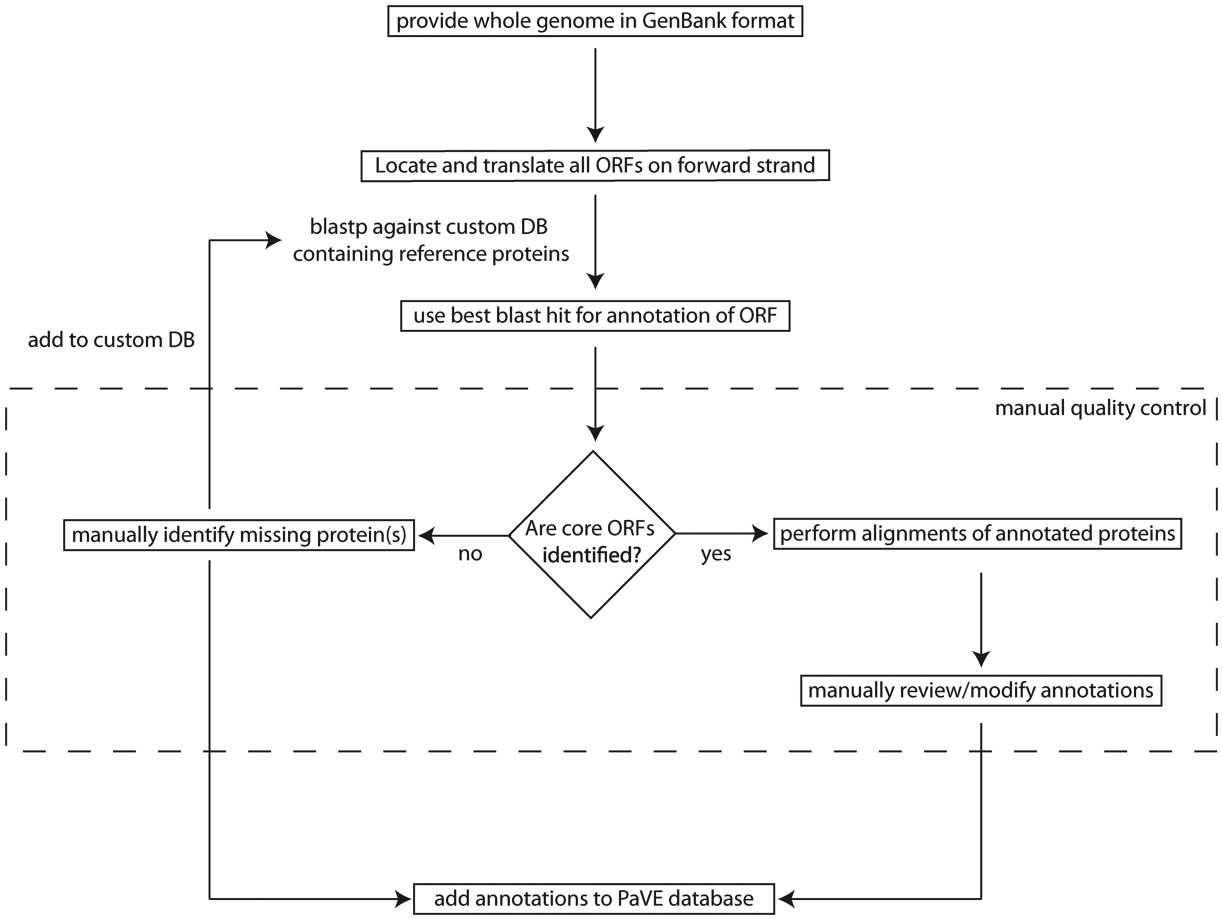

Figure 2. Genome annotation pipeline employed in the PaVE: the flowchart outlines the different steps used to curate the annotation of viral ORFs. See the main text for a detailed description.

the GenBank display format of each genome sequence on the PaVE website.

Manual editing of viral genomes

In 1997 John Meissner noted that: 'Fifteen years ago, sequencing a papillomavirus genome was both a significant undertaking and a remarkable achievement. Reagents and enzymes did not come as numbered tubes in quality-controlled kits. Equipment was basic. Techniques were still being worked out. That these initial sequence determinations contained as few errors as they did—produced under the added pressure of finishing before competing laboratories—is a testament to the scientific abilities of the researchers. Without question, these pioneering efforts should be judged by a different standard. Without resequencing, though, these uncorrected sequences will continue to generate unnecessary confusion.' (13). The PaVE subscribes to the same philosophy and has elected to edit viral genomes when sufficient evidence is available to do so. The Los Alamos HPV database group corrected the sequences for BPV1, HPV1, HPV5, HPV6, HPV16 and HPV18 (13) based on published evidence of sequencing errors and we have incorporated those corrected sequences into the PaVE database.

Several other viral genome sequences contained putative point mutations that interrupted well-characterized ORFs. Some of these apparent mutations could be due to sequencing errors, but in other cases the mutation may be present in the genome that was directly isolated from clinical material and may have contributed to disease outcome. However, for the purposes of comparative genomics, it is essential to have full-length protein sequences. Furthermore, while clinical lesions may contain such viruses, it is unlikely that these viral genomes could efficiently replicate and be propagated as circulating viruses. Thus, in these cases, we use information derived from other very closely related isolates to support our edits and 'repair' the genome. Both original and corrected genomes are available in the PaVE, as are pairwise alignments between the original and edited sequences. Table 1 summarizes the edits incorporated into the PaVE database, to date.

Searching capabilities

Users can query the PaVE database in three main ways: (i) a simple keyword search; (ii) advanced queries constructed to include host species, taxonomic classification, genes or regions and (iii) BLAST-based similarity search. The results of each query are dynamically updated and can

Downloaded from http://nar.oxfordjournals.org/ by guest on December 9, 2015

Table 1. Differences between PaVE Reference clones and GenBank sequences

Virus	NCBI Number	Differences between PaVE Reference clones and GenBank sequences									
HPV15	X74468	2809 E2 (G>-)									
HPV82	AB027021	5508 L2 (->C)									
CgPV1	GU014532	2641 E1 (->A)									
CPV11	JF800658	5757 L2 (->C)									
HPV72	X94164	2145 E1 (G>-)									
EcPV2	EU503122	1093 E1 (G>-)									
HPV6	X00203	7351 (insertion of[a])									
HPV14	X74467	628 (insertion of[b])									
HPV56	X74483	1091 E1 (C>-)	2810 E2 (->A)	3821 E2 (C>-)							
HPV53	X74482	1102 E1 (C>-)	1269 E1 (A>-)	1588 E1 (A>-)							
BPV1	X02346	1205 E1,E8 (T>N)	3445 E2 (G>-)	7306 URR (C>G)	7589 URR (->G)	7763 URR (C>-)					
HPV5	M17463	4055 E2 (TGCT>AATG)	6175 L1 (G>C)	6265 L1 (G>C)	6502 L1 (G>C)						
HPV1	V01116	1283 E1 (G>A)	2301 E1 (C>T)	3886 E5 (A>-)	4332 L2 (T>A)	4376 L2 (AG>GA)	4380 L2 (A>G)	4382 L2 (G>A)	7692 URR (GG>CC)		
HPV18	X05015	287 E6 (G>C)	2856 E2 (GCGT> TGCG)	3084 E2 (GC>CG)	3275 E2 (G>C)	5701 L1 (G>C)	6460 L1 (G>C)	6625 L1 (G>C)	6842 L1 (G>C)		
HPV16	K02718	1139 E1 (G>-)	2926 E2 (G>A)	3907 E5 (T>-)	4365 L2 (T>A)	6242 L1 (G>C)	6903 L1 (CAT>—)	6956 L1 (—>GAT)	7437 URR (C>-)	7439 URR (G>C)	7867 URR (->A)
HPV71	AB040456	226 E6 (AT>—)	257 E6 (T>-)	575 E7 (G>-)	5503 L2 (G>-)	5581 L2 (->C)	5771 L2 (->G)	5812 L2 (G>-)	5890 L1(->C)	5900 L1 (G>-)	6537 L1 (C>-)
		6547 L1 (G>-)	6568 L1 (T>-)	6586 L1 (TTC>—)	6605 L1 (A>-)	6612 L1 (T>-)	6623 L1 (T>-)	6632 L1 (G>-)	6642 L1 (->C)	6648 L1 (GAA>—)	6679 L1 (G>-)
		6826 L1 (G>-)	6852 L1 (G>-)	7602 URR (T>-)							

A pairwise alignment between the PaVE RefClone and sequence available on GenBank was created. For each single nucleotide polymorphism, the position in the pairwise alignment is given, followed by the nucleotide change at this position (PaVE → GenBank). Finally, the affected ORF is indicated. HPV6 and HPV14 have large insertions, the sequence of these insertions is provided below.

[a]ATGTACTGTTATATGTATGTGTGTTGTATATATGTGTGTATATATGTGTCTGTGTGTATATGTATATGTATGTGTTGTGTATATATATGTGTGT.

[b]GACATCTGTGGCGAAAGCTTCCCTTTCATAAAGTGAGAGGCTCTTGGAAAGGAATCTGTAGGCTGTGTAAGCATTTTCAATATGATTGGTAAAGAGGTCACATTGCAAGAT
ATTGTTCTGGAGTTGAATGAATTGCAGCCAGAGGTACAACCAGTTGACCTGTTTTGTGAAGAGGAGTTACCGAATGAGCTGCAGGAAACAGAGGTGGAGCTTCATATCGAG
AGGACCGCGTACAAAGTTGTTGTACCTTGCGGCTGCTGCAAGGTTAAGCTT.

Downloaded from http://nar.oxfordjournals.org/ by guest on December 9, 2015

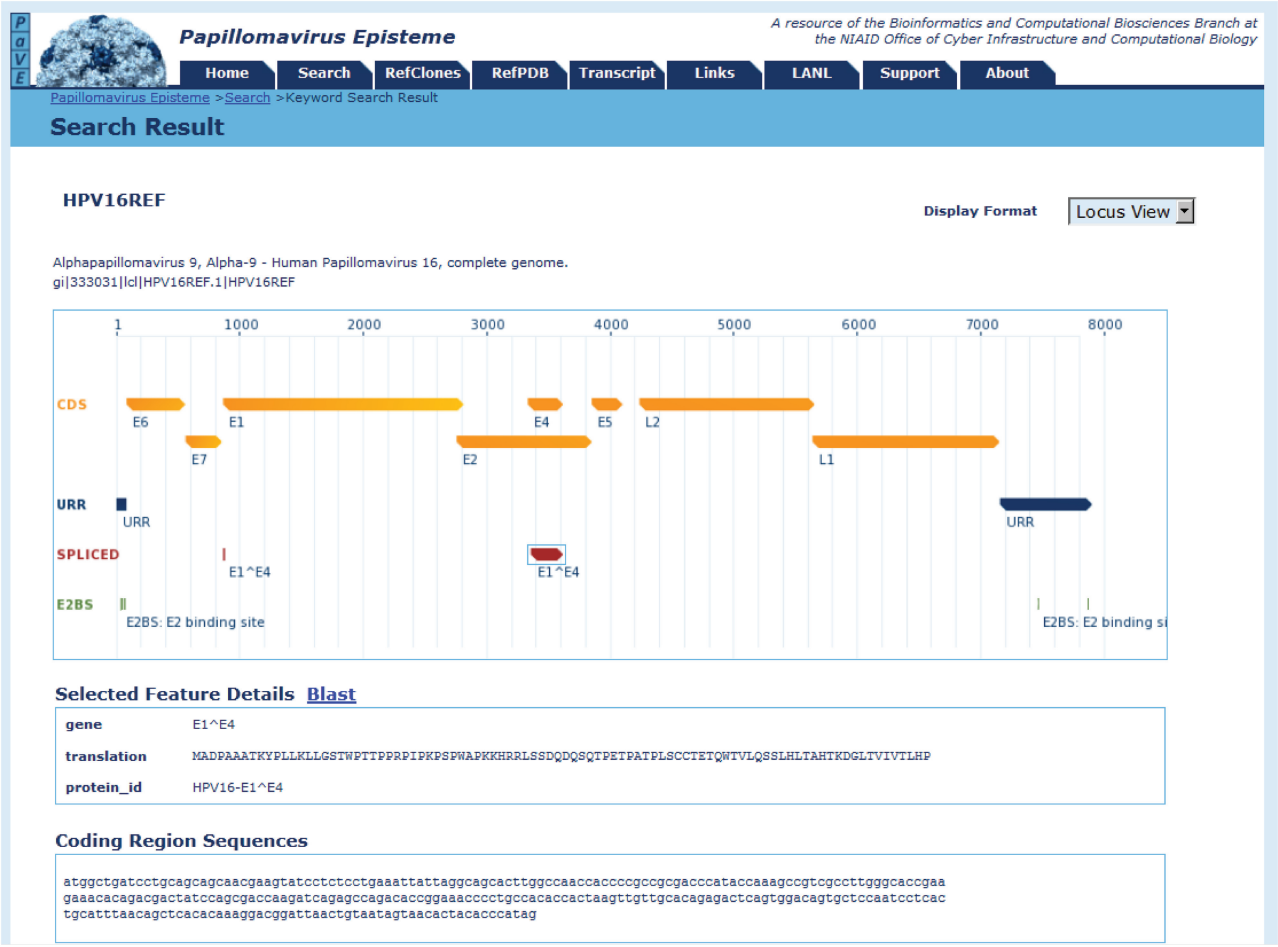

Figure 3. The PaVE locus viewer. The PaVE locus view shows a linear representation of the HPV16REF genome. Different annotations are displayed and can be selected. The annotated features include ORFs, spliced proteins, protein-binding site motifs and the viral URR. Upon selection, additional information is displayed in the 'selected feature details' window. The 'coding region sequences' window shows the nucleotide sequence for the selected ORF. Protein sequences can be directly compared with others in the PaVE database using the provided link to BLAST.

be downloaded for further analyses. Selecting genomes, genomic regions or proteins from any of the virus types will direct the user to the locus viewer for further information (Figure 3).

Browse Reference genomes

The 'RefClone' section of the website provides users with an agnostic entry-point into the PaVE database. Users can access viral types through an interactive phylogenetic tree or can browse tables of Animal or Human viral types to access additional details about each virus (e.g. common name of the host, associated publications, GenBank accession numbers). These tables can be downloaded as comma separated (csv) text files. Selecting a virus type in the phylogenetic tree or tables will direct the user to the locus viewer (Figure 3).

The PaVE locus viewer

The locus viewer displays a linear representation of the viral genome with annotated features. Each feature can be selected, and details for this feature will be displayed in a panel below. In addition, the selected feature is highlighted in the nucleotide sequence of the genome. From here, the information associated with the feature can be copied, downloaded or compared with other features in the PaVE database using BLAST.

Structure viewer

All experimentally derived PV protein structures available from the PDB (http://www.rcsb.org) are integrated into the PaVE database. In addition to viewing and analyzing these protein structures (visualized using JMol (v.11.9.6), the PaVE Structure Viewer computes and displays a pairwise sequence comparison between the solved structure and any homologous protein in the PaVE database. Any of these homologous proteins (BlastP E $<$ 1e$-$5) can be readily selected (Figure 4) and the alignment displayed. The user can opt to color code differences and identities between the sequences and can select portions of the sequence alignment to be highlighted on the structure.

PLANS FOR FUTURE EXPANSION

We are hoping to expand the PaVE database to include additional genome sequence and disease-related

Downloaded from http://nar.oxfordjournals.org/ by guest on December 9, 2015

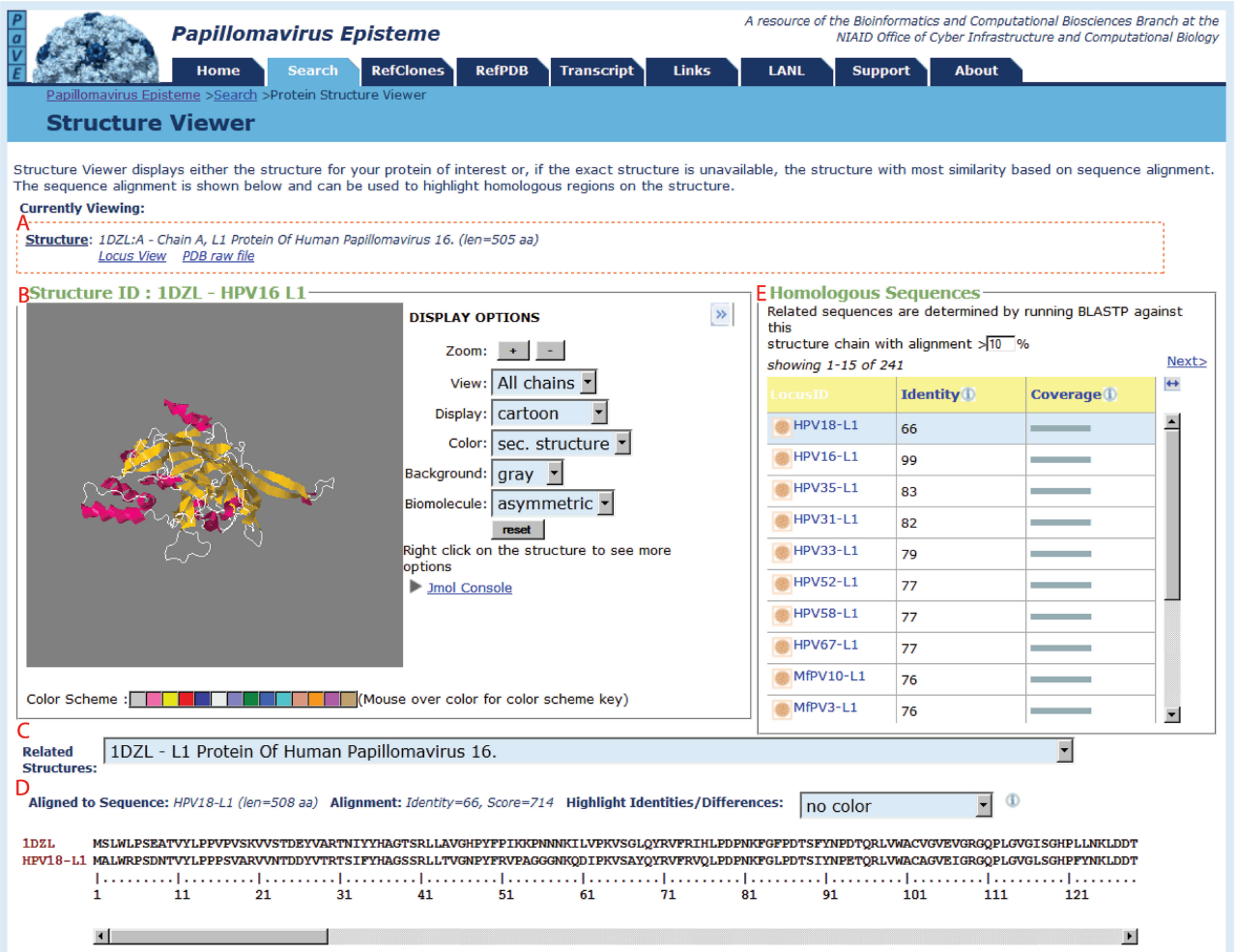

Figure 4. The PaVE structure viewer. The HPV16 L1 structure (PDB No. 1DZL) is shown within the PaVE structure viewer. The PaVE structure viewer consists of five different modules. Module (**A**) shows basic information about the structure under investigation and includes links to the locus viewer and the original PDB file. Module (**B**) is constructed around a fully functional Jmol Console (www.Jmol.org), allowing manipulation of the viral structure. For some viral proteins, the structures have been solved for several homologous proteins isolated from different viral types. These alternative structures can be selected using a pull down menu in (**C**). A unique feature of the PaVE structure viewer is the pairwise alignment shown in (**D**). Under default conditions, this alignment shows the sequence of the PaVE RefClone protein to the sequence present in the PDB file. However, users can use module (**E**) to compare homologous sequences with the displayed structure. Statistics derived from the BLASTp alignment are provided and differences or identities can be highlighted on the alignment. Specific residues or groups of residues can be selected and highlighted on the structure.

information, new tools and expert opinions on many aspects of PV biology.

An image library

PVs infect a wide variety of hosts including snakes, birds, sea and land mammals. Within each host they infect specific ecological niches on the mucosal and cutaneous epithelium, which results in a wide range of clinical outcomes. PV infection can result in flat macules or papular warts, endophytic or exophytic verrucas and condylomas that can present as small papules or large tumor masses. Many infections are sub-clinical in immune-competent hosts, but give rise to diverse and extensive lesions when the immune system is compromised. In the future, we hope to incorporate an image library into the PaVE database to provide users with gross clinical and histopathological images from a variety of lesions associated with each PV type. The powerful, dynamic search engine of PaVE will allow users to filter images of interest. Some authors (K.V.D. and A.A.M. are in the process of acquiring images from clinicians and researchers throughout the world and hopefully the PaVE will host some historical image collections that would otherwise be lost.

Bioinformatics focused review chapters by PV experts

To complement the PaVE website, one of us (A.A.M.) is co-editing a Special Open Access Issue of the Elsevier journal 'Virology'. This Special Issue will contain a collection of reviews written by leaders in the field that focus on structural and functional analysis of each viral protein or regulatory region; discuss epigenetic regulation and describe viral transcripts, viral genome variants, genome classification and viral evolution. Once published, we plan to link to these chapters directly from the PaVE website to

Downloaded from http://nar.oxfordjournals.org/ by guest on December 9, 2015

provide the user with state of the art knowledge about viral proteins and features. In turn, the PaVE will be able to use the expert knowledge in these chapters to refine and expand genome and protein features annotated in the PaVE.

Expanded annotation of genome and protein features

The viral genomes are rich in features of interest to researchers such as transcriptional promoters, enhancers and other regulatory elements, mRNA splice sites, replication elements and binding sites for cellular and viral proteins. PV proteins are well studied and much is known about functional domains, sites of protein–protein interaction and post-translational modifications. The PaVE Special Issue in Virology will contain detailed information and analyses of these features by experts in the field, which will enable us to develop a more detailed annotation pipeline. In the future, these features could be incorporated into the locus viewer and be available as filters on the search page of PaVE.

Multiple sequence alignments and phylogenetic trees

Sequence alignments of homologous proteins from different PV types are one of the most useful and informative techniques for comparative functional analyses as well as to study the evolutionary history and relationship among different PVs. In the next phase of PaVE, we hope to incorporate tools to allow users to generate multiple sequence alignments and to build phylogenetic trees.

PV genome variants

At present the PaVE hosts reference sequences for each viral type that has been completely cloned and sequenced. However, many variant sequences also exist for each viral type. There is a great interest among clinicians, epidemiologists and basic researchers in determining whether specific variants are associated with specific disease outcomes such as progression to cancer. Our future plans are to include databases of variant genomic sequences that are integrated with existing PaVE tools.

PV typing tool

The taxonomic classification of PVs has traditionally been based on nucleotide identity across the L1 ORF. For a viral isolate to be considered a novel type, it cannot share >90% sequence identity with any other known PV (1). An L1-specific typing tool has been developed by Piet Maes (Katholieke Universiteit Leuven) based on a similar application developed for rotavirus classification (14). The PV-based tool will compare the user's L1 sequence with the L1 sequences from all named PV types (a beta-version of this application is currently available at http://www.regatools.be/pavic10/). Users are informed whether their new isolate is different from other named viruses and meets the criteria for a new type. We hope to incorporate this tool into the PaVE site and provide users with information on how to submit their genomic information to the Papillomavirus Reference Centers (1,2). It is important to note that only the Reference Centers (together with the International Committee on the Taxonomy of Viruses) have the authority to name viral types and the L1 typing tool is intended only to aid the researcher with the initial steps in submitting a putative novel viral type.

CONCLUSIONS

In the few years since its initial release, the PaVE has established itself as a centralized body of PV sequence-related information. The PaVE currently hosts 241 completely sequenced and annotated Reference Viral genomes. The data on the PaVE can be downloaded for 'local' analyses but an ever-expanding array of bioinformatics tools is provided to aid the user in analyzing their data online.

The relatively simple genomic organization and evolutionary history of PVs, combined with our well-curated collection of genome sequences, will be welcomed by evolutionary biologists and bioinformaticians outside of the PV field. The data on the PaVE will also be of interest to clinicians, epidemiologists and basic scientists alike. In the era of high-throughput sequencing, comparative genomics and epidemiological analyses of variant genomes, it is paramount that scientists have access to uniformly curated Reference genomes thereby ensuring that scientists across several disciplines study the same prototype genome (15). The goal of the PaVE is to seamlessly integrate carefully curated data with novel analytical tools that will assist in the study of PV biology and aid in the development of improved therapeutics and diagnostics.

ACKNOWLEDGEMENTS

The authors thank the PaVE advisors Hans-Ulrich Bernard, Thomas Brettin, Thomas Broker, Christopher B. Buck, Robert D. Burk, John Doorbar, Ethel-Michele de Villiers and Marc Van Ranst. They are also grateful to members of the McBride Laboratory for testing the PaVE. They thank Yongjian Guo for his contribution to the technical design and implementation of the PaVE and Jason Barnett for his contribution to the PaVE user-interface design.

FUNDING

The Intramural Research Program [1ZIAAI001071] and Office of Science Management and Operations of the National Institutes of Allergy and Infectious Diseases, National Institutes of Health. Funding for open access charge: NIAID Intramural Research Program [1ZIAAI001071].

Conflict of interest statement. None declared.

REFERENCES

1. de Villiers,E.M., Fauquet,C., Broker,T.R., Bernard,H.U. and zur Hausen,H. (2004) Classification of papillomaviruses. *Virology*, **324**, 17–27.
2. Bernard,H.U., Burk,R.D., Chen,Z., Van Doorslaer,K., zur Hausen,H. and de Villiers,E.M. (2010) Classification of papillomaviruses (PVs) based on 189 PV types and proposal of taxonomic amendments. *Virology*, **401**, 70–79.

Downloaded from http://nar.oxfordjournals.org/ by guest on December 9, 2015

3. Trottier,H. and Franco,E.L. (2006) The epidemiology of genital human papillomavirus infection. *vaccine*, **24 (Suppl. 1)**, S1–15.
4. Bouvard,V., Baan,R., Straif,K., Grosse,Y., Secretan,B., El,G.F., Benbrahim-Tallaa,L., Guha,N., Freeman,C., Galichet,L. *et al.* (2009) A review of human carcinogens—Part B: biological agents. *Lancet Oncol.*, **10**, 321–322.
5. Garcia-Vallve,S., Alonso,A. and Bravo,I.G. (2005) Papillomaviruses: different genes have different histories. *Trends Microbiol.*, **13**, 514–521.
6. Klingelhutz,A.J. and Roman,A. (2012) Cellular transformation by human papillomaviruses: lessons learned by comparing high- and low-risk viruses. *Virology*, **424**, 77–98.
7. Rector,A., Lemey,P., Tachezy,R., Mostmans,S., Ghim,S.J., Van Doorslaer,K., Roelke,M., Bush,M., Montali,R.J., Joslin,J. *et al.* (2007) Ancient papillomavirus-host co-speciation in Felidae. *Genome Biol.*, **8**, R57.
8. Rector,A., Stevens,H., Lacave,G., Lemey,P., Mostmans,S., Salbany,A., Vos,M., Van Doorslaer,K., Ghim,S.J., Rehtanz,M. *et al.* (2008) Genomic characterization of novel dolphin papillomaviruses provides indications for recombination within the Papillomaviridae. *Virology*, **378**, 151–161.
9. Ho,L., Chan,S.Y., Burk,R.D., Das,B.C., Fujinaga,K., Icenogle,J.P., Kahn,T., Kiviat,N., Lancaster,W., Mavromara-Nazos,P. *et al.* (1993) The genetic drift of human papillomavirus type 16 is a means of reconstructing prehistoric viral spread and the movement of ancient human populations. *J. Virol.*, **67**, 6413–6423.
10. Bao,Y., Federhen,S., Leipe,D., Pham,V., Resenchuk,S., Rozanov,M., Tatusova,R. and Tatusova,T. (2004) National center for biotechnology information viral genomes project. *J Virol.*, **78**, 7291–7298.
11. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
12. Katoh,K., Misawa,K., Kuma,K. and Miyata,T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, **30**, 3059–3066.
13. Meissner,J.D. (1997) Sequencing errors in reference HPV clones. In: Myers,G., Baker,C., Munger,K., Sverdrup,F., McBride,A. and Bernard,H.-U. (eds), *Human Papillomaviruses 1997*. Los Alamos National Laboratory, Los Alamos, NM, pp. III110–III123.
14. Maes,P., Matthijnssens,J., Rahman,M. and Van Ranst,M. (2009) RotaC: a web-based tool for the complete genome classification of group A rotaviruses. *BMC Microbiol.*, **9**, 238.
15. Wilkinson,D.E., Baylis,S.A., Padley,D., Heath,A.B., Ferguson,M., Pagliusi,S.R., Quint,W.G. and Wheeler,C.M. (2010) Establishment of the 1st World Health Organization international standards for human papillomavirus type 16 DNA and type 18 DNA. *Int. J. Cancer*, **126**, 2969–2983.

Downloaded from http://nar.oxfordjournals.org/ by guest on December 9, 2015